How to beat the sphere-packing bound with feedback

Anant Sahai Wireless Foundations, Dept. of EECS
University of California at Berkeley
Email: sahai@eecs.berkeley.edu

Abstract

The sphere-packing bound $E_{sp}(R)$ bounds the reliability function for fixed-length block-codes. For symmetric channels, it remains a valid bound even when strictly causal noiseless feedback is allowed from the decoder to the encoder. To beat the bound, the problem must be changed. While it has long been known that variable-length block codes can do better when trading-off error probability with expected block-length, this correspondence shows that the *fixed-delay* setting also presents such an opportunity for generic channels.

While $E_{sp}(R)$ continues to bound the tradeoff between bit error and fixed end-to-end latency for symmetric channels used *without* feedback, a new bound called the "focusing bound" gives the limits on what can be done with feedback. If low-rate reliable flow-control is free (ie. the noisy channel has strictly positive zero-error capacity), then the focusing bound can be asymptotically achieved. Even when the channel has no zero-error capacity, it is possible to substantially beat the sphere-packing bound by synthesizing an appropriately reliable channel to carry the flow-control information.

How to beat the sphere-packing bound with feedback

I. INTRODUCTION

The two most fundamental parameters when it comes to reliable data transport are end-to-end system delay and the probability of error. Error probability is fundamental because a low probability of bit error lies at the heart of the digital revolution justified by the source/channel separation theorem. Delay is important because it is the most basic cost that a system must pay in exchange for reliability — it allows the laws of large numbers to be harnessed to smooth out the variability introduced by random communication channels.

Traditionally, block-length has been used as a proxy for end-to-end delay since block-codes are easier to understand than non-block codes. Even when fixed end-to-end delay is desired, this paper shows that nonblock codes can provide a tremendous advantage when feedback is allowed. This short correspondence is a companion to our longer work in [1]. Some key results are reviewed here in the next section, but the reader is referred to [1] for more details, motivation, as well as a perspective on the existing results in the literature. The new contribution in this correspondence comes in Section III. It shows how to construct a special fixed-delay code over a DMC using noiseless feedback. It beats the sphere-packing bound with fixed-delay in the high-rate regime even for channels (like the BSC) that have no zero-error capacity. A plot is given for the BSC-0.4 case that provides an explicit counterexample to Pinsker's assertion (Theorem 8 in [2]) that this is impossible to do.

Simsek had earlier built codes for the BSC in [3], [4] which beat the sphere-packing bound with fixed delay. Those were fundamentally built upon the equivalence between scalar stabilization problems and feedback communication problems established in [5], [6], but were hard to analyze. They also did not work at high rates. The advantage of the codes given here is their conceptual simplicity and the fact that they beat the sphere-packing bound in the high-rate regime. These codes do not do well in the low-rate regime and it is clear that they could be married with Simsek-codes to give some improvements at low-rate. However, this would not be enough to reach the focusing bound so there is much room for improvement.

II. REVIEW

A. Block coding

The fundamental lower-bound on error probability comes from the sphere-packing or volume bound, and this bound is also known to be achievable at high rates by random-coding [7]. Reliable communication is not possible if during the block, the channel acts like one whose capacity is less than the target rate. Following [8] and [9], for block codes this idea immediately gives the following bound on the exponential error probability:

$$E^+(R) = \inf_{G:C(G)<R} \max_{\vec{r}} D\left(G||P|\vec{r}\right) \tag{1}$$

where $D(G||P|\vec{r})$ is the divergence term that governs the exponentially small probability of the true channel P behaving like channel G when facing the input distribution coming from the codeword composition $\vec{r}$.

Even with causal noiseless feedback, there is no way around this bound because channel capacity does not increase with feedback for memoryless channels. Without feedback, the bound can be tightened to the form traditionally known as the sphere-packing bound.

$$E_{sp}(R) = \max_{\vec{r}} \min_{G:I(\vec{r},G)\leq R} D\left(G||P|\vec{r}\right) \tag{2}$$

For symmetric channels, the optimizing codeword composition $\vec{r}$ is always uniform and $E_{sp}(R) = E^+(R)$. Thus, for fixed-block codes and symmetric DMCs, no only does causal feedback not improve capacity, but it does not improve reliability either[10]!

An alternate form for $E_{sp}(R)$ is given by:

$$E_{sp}(R) = \max_{\rho>0} \left[E_0(\rho) - \rho R\right] \tag{3}$$

with the Gallager function $E_0(\rho)$ defined as:

$$E_0(\rho) = \max_{\vec{q}} -\ln \sum_y \left[\sum_x q_x p_{x,y}^{\frac{1}{1+\rho}}\right]^{(1+\rho)} \tag{4}$$

Note that for symmetric channels, it suffices to use a uniform $\vec{q}$ while optimizing (4). Also, since the random-coding error exponent is given by:

$$E_r(R) = \max_{0<\rho\leq1} \left[E_0(\rho) - \rho R\right] \tag{5}$$

It is clear that the sphere-packing bound is achievable, even without feedback, at rates close to C since for those rates, $\rho < 1$ optimizes both expressions [7]. The points on the sphere-packing bound where $\rho > 1$ are also achievable by random coding if the sense of "correct decoding" is slightly relaxed. Rather than forcing the decoder to emit a single estimated codeword, list-decoding allows the decoder to emit a list of guessed codewords. The decoding is considered correct if the true codeword is on the list. For list-decoding with list size ℓ in the context of random codes, Problem 5.20 in [7] reveals that

$$E_{r,\ell}(R) = \max_{0<\rho\leq\ell} \left[E_0(\rho) - \rho R\right] \tag{6}$$

is achievable. At high rates (where the maximizing ρ is small), there is no benefit from relaxing to list-decoding, but it makes a difference at low rates.

B. Non-block codes

Another classical approach to the problem of reliable communication is to consider codes without a block structure. Convolutional and tree codes represent the prototypical examples. It was realized early on that in an infinite[1] constraint length convolutional code under ML decoding, all bits will eventually be decoded correctly [7]. However, if the end-to-end delay is forced to be bounded, then the bit error probability with delay is governed by $E_r(R)$ for random convolutional codes, even when the constraint lengths are unbounded [11]. This performance with delay is also achievable using an appropriately biased sequential decoder [12]. A nice feature of sequential decoders is that they are not tuned to any target delay — they can be prompted for estimates at any time and they will give the best estimate that they have. Thus an infinite constraint-length convolutional code with appropriate sequential decoding achieves the exponent $E_r(R)$ delay universally over all (sufficiently long) delays.

Pinsker claimed in [2] that the sphere-packing bound continued to bound the performance of nonblock codes both with and without feedback. He had proofs for the BSC case, but asserted that the result held more generally. While he was right for the without feedback case, it turns out that there is a subtle flaw in his argument regarding the case with feedback.

1) The BEC example: This example, repeated from [1] for the reviewer's convenience, shows the power of feedback in the delay context. The binary erasure channel with erasure probability $\delta < \frac{1}{2}$ used at bit-rate $R' = \frac{1}{2}$ gives a counterexample to Pinsker's conjecture. The BEC is so simple that everything can be understood with a minimum of overhead.

$$E_{sp}(\frac{1}{2}) = D(\frac{1}{2}||\delta) = -\frac{\ln(4\delta(1-\delta))}{2} \tag{7}$$

For $\delta = 0.4$, this corresponds to an error exponent of about 0.02. Even with feedback, there is no way for a fixed block-length code to beat this exponent. If the channel lets fewer than $\frac{n}{2}$ bits through the channel, it is impossible to reliably communicate an $\frac{n}{2}$ bit message!

---

[1]More precisely, these are unbounded constraint length codes since at any finite time there are only a finite number of data bits so far.
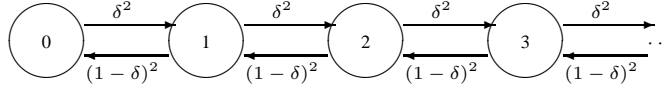
Fig. 1. The birth-death Markov chain governing the rate $\frac{1}{2}$ feedback communication system over an erasure channel.

If causal noiseless feedback is available, the natural nonblock code just retransmits a bit until it is correctly received. As bits arrive steadily at the rate $R' = \frac{1}{2}$, they enter a FIFO queue of bits awaiting transmission. If we look at the queue state every two channel uses, it can be modeled (see Figure 1) as a birth-death Markov chain with a δ^2 probability of birth and a $(1 - \delta)^2$ probability of death. Converting that into an error exponent with delay d gives:

$$E_f^{bec}(\frac{1}{2}) = \ln(1 - \delta) - \ln(\delta) \tag{8}$$

Plugging in $\delta = 0.4$ gives an exponent of more than 0.40. This is about twenty times higher than the sphere-packing bound!

C. The focusing bound

Restricting attention to symmetric channels, the BEC case can be abstracted to get a general bound on the probability of error with delay. [1] calls this bound the "focusing bound" because it is based on the idea of having the encoder focus as much of the decoder's uncertainty as possible onto bits whose deadlines are not pending.

Definition 2.1: A *rate R encoder with noiseless feedback* is a sequence of maps $\mathcal{E}_t$. The range of each map is the discrete set $\mathcal{X}$. The t-th map takes as input the available data bits $B_1^{\lfloor R't \rfloor}$, as well as all the past channel outputs Y_1^{t-1}.

Randomized encoders with noiseless feedback also have access to a continuous uniform random variable W_t denoting the common randomness available in the system.

Definition 2.2: A *delay d rate R decoder* is a sequence of maps $\mathcal{D}_i$. The range of each map is just an estimate $\widehat{B}_i$ for the i-th bit taken from $\{0, 1\}$. The i-th map takes as input the available channel outputs $Y_1^{\lceil \frac{i}{R'} \rceil + d}$ which means that it can see d time units beyond when the bit to be estimated first had a chance to impact the channel inputs.

Randomized decoders also have access to all the continuous uniform random variables W_t.

Definition 2.3: The fixed-delay error exponent α is asymptotically *achievable* at rate R across a noisy channel if for every delay d_j in some increasing sequence $d_j \to \infty$ there exist rate R encoders and delay d_j decoders $\mathcal{E}^{d_j}, \mathcal{D}^{d_j}$ that satisfy the following properties when used with input bits B_i drawn from iid fair coin tosses.

1) For every j, there exists an $\epsilon_j < 1$ so that $P(B_i \neq \widehat{B}_i(d_j)) \leq \epsilon_j$ for every $i \geq 1$. The $\widehat{B}_i(d_j)$ represents the delay d_j estimate of B_i produced by the $\mathcal{E}^j, \mathcal{D}^j$ pair connected to the input B and the channel in question.
2) $\lim_{j \to \infty} \frac{-\ln \epsilon_j}{d_j} \leq \alpha$

The exponent α is asymptotically *achievable universally over delay* or in an *anytime fashion* if a single encoder $\mathcal{E}$ can be used above for all d_j above.

Theorem 2.4: Focusing bound from [1]: For a discrete memoryless channel, no delay exponent $\alpha > E_a(R)$ is asymptotically achievable even if the encoders are allowed access to noiseless feedback.

$$E_a(R) = \inf_{0 < \lambda < 1} \frac{E^+(\lambda R)}{1 - \lambda} \tag{9}$$

where E^+ is the Haroutunian exponent from (1). When the DMC is symmetric, $E_a(R)$ can be expressed parametrically as:

$$E_a(R) = E_0(\eta) \; ; \; R = \frac{E_0(\eta)}{\eta} \tag{10}$$

where $E_0(\eta)$ is the Gallager function from (4), and η ranges from 0 to ∞.

D. The (n, c, l) family of codes

The focusing bound is attained for the BEC with feedback using the natural "repeat bits until successful" code. As demonstrated in [1], it can also be asymptotically attained for any noisy channel provided we have access to a low-rate channel that can deliver perfectly noiseless flow-control bits from the encoder to the decoder. The code is reviewed below.

Call $c \geq 1$ the chunk length, 2^l the list length, and $n > l$ the data block length. The (n, c, l) scheme is:

- Queue up incoming bits and assemble them into blocks of size $\frac{ncR}{\ln 2}$ bits. If there are fewer than $\frac{ncR}{\ln 2}$ bits still awaiting transmission, just idle by transmitting an arbitrary input letter.
- At every noisy channel use, the encoder sends the channel input corresponding to the next position in an infinite-length random codeword associated with the current data block, where the random codewords are drawn iid using the appropriate input distribution[2] over the noisy channel's input alphabet.
- If the time is an integer multiple of c, use the noiselessly fedback channel outputs to simulate the decoder's attempt to decode the current codeword to within a list of the top 2^l items. If the true data-block is one of the 2^l items, send a 1 over the noiseless flow-control link. Also send the disambiguating l bits representing the true block's index within the decoder's list. Remove the current block of $\frac{ncR}{\ln 2}$ bits from the main data queue as well. If the true block is not in the decoder's list, just send a 0 over the noiseless flow-control link.
- At the decoder, the encoder queue length is known perfectly since it can only change by the deterministic arrival of data bits or when a noise-free confirm or deny bit has been sent over the flow-control link. Thus the decoder always knows which input block a given channel output Y_t or fortified symbol S_t corresponds to.
- If the time is an integer multiple of c and the decoder receives a 1 noiselessly, then it decodes what it has seen to a list of the top 2^l possibilities for this block. It will use the next l noisefree flow-control bits to disambiguate this list and will use the result as its estimate for the block.

Such schemes are shown in [1] to be asymptotically optimal:

Theorem 2.5: By appropriate choice of (n, c, l), it is possible to asymptotically achieve all delay exponents $\alpha < E_0(\rho)$ for $R = \frac{E_0(\rho)}{\rho}$ for the fortified system built around a DMC by adding a rate $\frac{1}{k}$ noisefree forward flow-control link where k can be made as small as desired.

III. SYNTHESIZING A PATHWAY TO CARRY FLOW-CONTROL INFORMATION

These codes use time-sharing of the channel to split it into two parts. One part carries the data and the other part carries flow control information.

A. Channels with positive zero-error capacity

The fortified communication scheme is easily adapted to channels with strictly positive zero-error capacity by just using the feedback zero-error capacity to carry the flow-control information [1]. There is no k. Instead, let θ be block-length required to realize feedback zero-error transmission of at least $l + 1$ bits. As illustrated in Figure 2, terminate each chunk with a length θ feedback zero-error code and use it

---

[2]Use the $E_0(\eta)$ maximizing input distribution for the η such that the data rate $R = \frac{E_0(\eta)}{\eta}$.
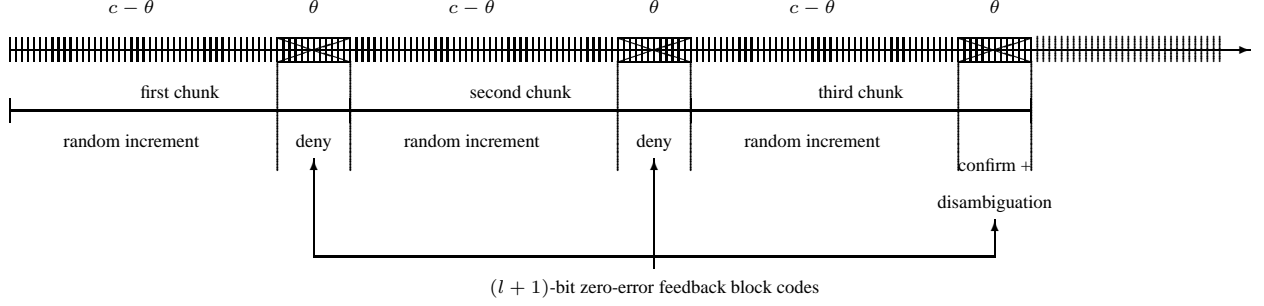
DMC channel uses

Fig. 2. One block's transmission in the (n, c, l, θ) code for noisy channels. The θ is used to carry $l + 1$ flow-control bits reliably. If the channel has a strictly positive feedback zero-error capacity, θ does not scale with c. If it does not, θ is proportional to c.

to transmit the flow-control information. If the chunk size is c, then it is as though we are operating with only a fraction $(1 - \frac{\theta}{c})$ of the channel uses. The overhead tends to zero by making the chunk sizes long giving the following corollary to Theorem 2.5:

Corollary 3.1: By appropriate choice of (n, c, l), it is possible to asymptotically achieve all delay exponents $\alpha < E_0(\rho)$ for $R = \frac{E_0(\rho)}{\rho}$ for any channel with $C_{0,f} > 0$.

B. Channels without zero-error capacity

When the channel has no zero error capacity, then we can still allocate θ channel uses per chunk to carry flow control information and have the encoder just assume that it was received correctly. This can be done by using an infinite constraint-length time-varying random convolutional code.[3] This gives a delay-universal scheme that is guaranteed to eventually get the flow-control information across correctly. Unlike a zero-error code, all that such a code can guarantee is that the probability of error in the entire message stream prefix is exponentially small in the number of channel uses that have occurred in the code since that message stream prefix was determined.

The flow-control information can be viewed as low-rate "punctuation" that tells the decoder how to parse the channel outputs that are carrying the data itself. Essentially, the punctuation gives "commas" that separate out the different message blocks[4]. Here, we assume that the decoder uses its current best estimate of the punctuation to re-parse the history of the data-carrying stream. Then the data-carrying channel outputs are decoded assuming that the flow-control information is correct. Any bits that have reached their deadlines are emitted, but this does not prevent the decoder from re-parsing them in the future.

Consequently, an error can occur at the decoder in two different ways. As before, the data-carrying stream could be corrupted due to channel atypicality in those slots. However, the flow-control stream could also become corrupted. As a result, the θ must be kept proportional to the chunk length c to avoid having the flow-control messages cause too many errors. The effective rate of the flow control information therefore goes to zero as $c \to \infty$ and the relevant error exponent is about $E_0(1)$. Balancing the error probabilities and optimizing over the choice of θ gives the following theorem:

Theorem 3.1: By appropriate choice of (n, c, l, θ), it is possible to asymptotically achieve all delay

---

[3]See [1] to see some tricks that allow such a code to be operated with feedback at bounded expected computational complexity, at least at low rates.

[4]As well as disambiguating any lists.

exponents $\alpha < E'(R)$ where the tradeoff curve is given parametrically by varying $\rho \in (0, \infty)$:

$$E'(\rho) = \left(\frac{1}{E_0(\rho)} + \frac{1}{E_0(1)} \right)^{-1} \tag{11}$$

$$R(\rho) = \frac{E'(\rho)}{\rho} \tag{12}$$

Proof: For simplicity of exposition, we assume that the block length n, chunk size c, and list size l are large enough that the code essentially achieves the focusing bound for whatever the effective rate is. The various ϵ terms are ignored.

Let ψ be the proportion of channel uses dedicated as overhead to run the low-rate flow-control channel. So the effective chunk size in the data-stream is $c' = c(1 - \psi)$. The effective rate of the message stream is thereby increased to $\frac{R}{1-\psi}$. Assuming that the flow control information is correct, the delay-universal error exponent is thus $E_a(\frac{R}{1-\psi})$ with respect to the delay in terms of code channel uses. But there are only $(1 - \psi)$ code channel uses per unit of actual time and so the delay exponent is $(1 - \psi)E_a(\frac{R}{1-\psi})$ with respect to true delay.

Meanwhile $\theta = c\psi$. The effective flow-control information rate is $\frac{l+1}{c\psi} \approx 0$ since c can be made as large as we want. Since this code achieves the random-coding error exponent, the delay-universal error exponent for the flow-control stream is essentially $E_0(1)$ with flow-code-channel uses since that is the zero-rate point for random coding. But there are only ψ flow-code-channel uses per actual time and so the delay-exponent for the flow-control stream is actually $\psi E_0(1)$ with respect to true delay.

Pick a fixed-delay d large. It can be written as $d = d_f + d_m$ in d different ways. Let d_f be the part of the delay that is "burned" by the flow-control stream. Thus, with probability exponentially small in d_f, this suffix of time has possibly incorrect flow-control information and so can not be trusted to be interpreted correctly. Thus, the performance of the code with delay d is like the performance of the underlying (n, c', l) code with delay d_m. Since the channel uses are disjoint, the two error events are independent and thus the achieved exponent is the weighted average of the two error exponents. Balancing the exponents of the two parts tells us to set:

$$E' = \psi E_0(1) = (1 - \psi)E_a(\frac{R}{1 - \psi})$$

with the resulting error probability with delay being governed by $\approx d \exp(-d\psi E_0(1))$. The polynomial term d in front is dominated entirely by the exponential decay and can be ignored.

Using the parametric forms using ρ for E_a, we get a pair of equations:

$$\psi E_0(1) = (1 - \psi)E_0(\rho) \tag{13}$$

$$\frac{E_0(\rho)}{\rho} = \frac{R}{1 - \psi} \tag{14}$$

The first thing to notice is that simple substitution gives

$$R = \frac{(1 - \psi)E_0(\rho)}{\rho} = \frac{\psi E_0(1)}{\rho} = \frac{E'}{\rho}$$

Solving for ψ shows (after a little algebra) that

$$\psi = \frac{E_0(\rho)}{E_0(1) + E_0(\rho)} \tag{15}$$

This way $1 - \psi = \frac{E_0(1)}{E_0(1) + E_0(\rho)}$ and the first equation is clearly true. Similarly $\frac{1}{1-\psi} = 1 + \frac{E_0(\rho)}{E_0(1)}$ and $\frac{\psi}{1-\psi} = \frac{E_0(\rho)}{E_0(1)}$ and thus the second equation is also true.
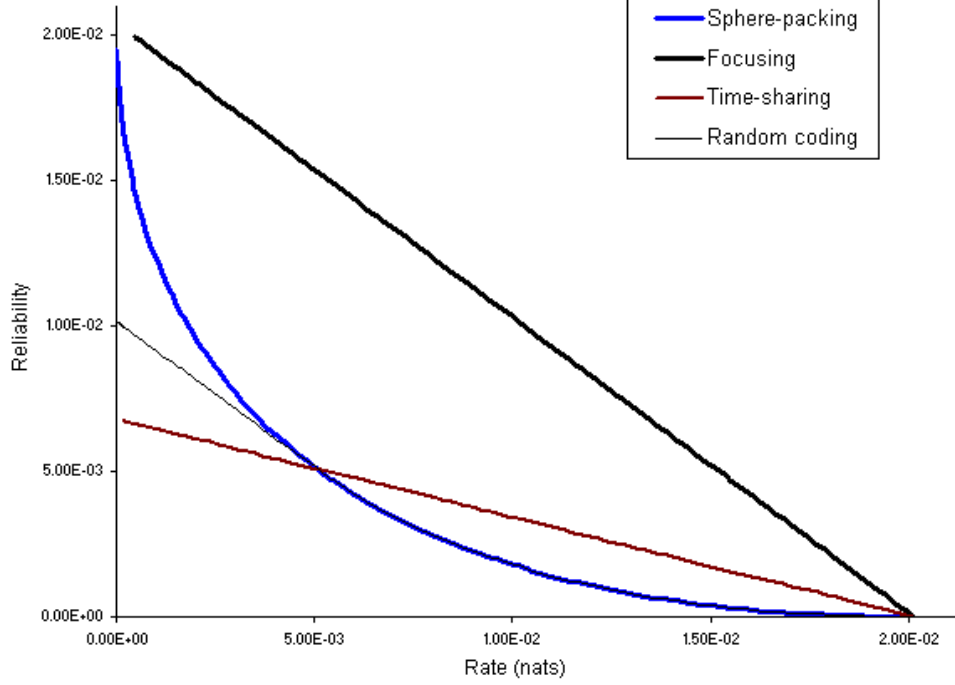
Fig. 3. The reliability functions for the binary symmetric channel with crossover probability $\delta = 0.4$. The sphere-packing bound approaches capacity quadratically flat while the focusing bound and the new scheme both approach the capacity point linearly.

$$
\begin{aligned}
E' &= \psi E_0(1) \\
&= \frac{E_0(\rho)E_0(1)}{E_0(1) + E_0(\rho)} \\
&= \left(\frac{1}{E_0(\rho)} + \frac{1}{E_0(1)} \right)^{-1}
\end{aligned}
$$

Which establishes the theorem. □

The superiority of these exponents to the sphere-packing bound in the high rate regime is immediately clear since they are basically like the focusing bound in form. Some algebra and simple calculus reveals that the focusing bound[5] has slope $2C/\frac{\partial^2 E_0(0)}{\partial \rho^2}$ in the vicinity of the $(C, 0)$ point, while the $E'(R)$ curve achieved by Theorem 3.1 has the lower slope $E_0(1)/(C - \frac{E_0(1)}{2C}\left(\frac{\partial^2 E_0(0)}{\partial \rho^2}\right))$. Either way, in generic cases, the reliability drops linearly in the neighborhood of capacity rather than in a quadratically flat manner.

Figure 3 illustrates the bounds for a BSC with crossover probability 0.4.

IV. CONCLUSIONS

Even when there is no zero error capacity, flow-control can be used to substantially beat the sphere-packing bound with respect to delay at high rates. The arguments from [1] dealing with fixed-delay feedback also apply to the new code and show that the reliabilities achieved here are still asymptotically achievable even if the feedback is delayed. The key is that our flow-control code does not need instantaneous feedback to achieve its internal reliability target $\approx E_0(1)$.

---

[5]When $\frac{\partial^2 E_0(0)}{\partial \rho^2} = 0$, page 143 in [7] reveals that the Sphere-packing bound is a straight line hitting zero at capacity. In such cases, the focusing-bound is bounded away from zero even in the neighborhood of capacity and hence this curve has an infinite slope.

We conjecture that the gap between the focusing bound and the reliabilities achieved by our scheme in the no-zero-error case is due to our *a-priori* splitting of the channel into dedicated data and flow-control links. The parallel channel coding advantage tells us that splitting a channel generally results in a loss of reliability. The codes in [3] performed much better in the low-rate regime because they had the flow-control information implicitly within the message stream itself.

ACKNOWLEDGMENTS

REFERENCES

[1] A. Sahai, "Why block length and delay are not the same thing," *IEEE Trans. Inform. Theory*, submitted. [Online]. Available: http://www.eecs.berkeley.edu/~\protect\kern+.1667em\relax$\protect\kern-.1667em\relax$sahai/Papers/FocusingBound.pdf

[2] M. S. Pinsker, "Bounds on the probability and of the number of correctable errors for nonblock codes," *Problemy Peredachi Informatsii*, vol. 3, no. 4, pp. 44–55, Oct./Dec. 1967.

[3] H. T. Simsek, "Anytime channel coding with feedback," Ph.D. dissertation, University of California, Berkeley, 2004.

[4] T. Simsek, R. Jain, and P. Varaiya, "Scalar estimation and control with noisy binary observations," *IEEE Trans. Automat. Contr.*, vol. 49, no. 9, pp. 1598–1603, Sept. 2004.

[5] A. Sahai, "Any-time information theory," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, 2001.

[6] A. Sahai and S. K. Mitter, "The necessity and sufficiency of anytime capacity for stabilization of a linear system over a noisy communication link. part I: scalar systems," *IEEE Trans. Inform. Theory*, Aug. 2006.

[7] R. G. Gallager, *Information Theory and Reliable Communication*. New York, NY: John Wiley, 1971.

[8] I. Csiszar and J. Korner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Akademiai Kiado, 1981.

[9] E. A. Haroutunian, "Lower bound for error probability in channels with feedback," *Problemy Peredachi Informatsii*, vol. 13, no. 2, pp. 36–44, 1977.

[10] R. L. Dobrushin, "An asymptotic bound for the probability error of information transmission through a channel without memory using the feedback," *Problemy Kibernetiki*, vol. 8, pp. 161–168, 1962.

[11] G. D. Forney, "Convolutional codes II. maximum-likelihood decoding," *Information and Control*, vol. 25, no. 3, pp. 222–266, July 1974.

[12] F. Jelinek, "Upper bounds on sequential decoding performance parameters," *IEEE Trans. Inform. Theory*, vol. 20, no. 2, pp. 227–239, Mar. 1974.